DOCUMENT RESUME

ED 208 039 TM 810 736

AUTHOR Petersen, Nancy S.; And Others
TITLE IRT Versus Conventional Equating Methods: A Comparative Study of Scale Stability.
PUB DATE Apr 81
NOTE 47p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Entrance Examinations; Comparative Analysis; *Equated Scores; *Latent Trait Theory; *Mathematical Models; Methods; Scoring; Test Reliability
IDENTIFIERS *Equipercentile Equating; *Linear Equating Method; Scholastic Aptitude Test

ABSTRACT

Three equating methods were compared in terms of magnitude of scale drift: equipercentile equating, linear equating, and item response theory (IRT) equating. A sample of approximately 2670 cases was selected for each pairing of a form of the Scholastic Aptitude Tests (SAT) and an anchor test. Of the two conventional equating methods, equipercentile equating was less satisfactory than linear equating. The IRT methods did not behave in the same manner for the verbal and the mathematical equating chains. For the verbal chain, the IRT methods were greatly superior to the conventional methods. For the mathematical chain, the Levine models of linear equating and the IRT concurrent method tended to give very similar results. If it were necessary to use the same equating method for both SAT-Verbal and SAT-Mathematical forms, it appears that the IRT concurrent method would be the best equating method for reducing scale drift over time. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED208039

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

IRT Versus Conventional Equating Methods: A Comparative Study of Scale Stability[1,2]

Nancy S. Petersen
Linda L. Cook
Martha L. Stocking

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. L. Cook

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 736

[1] A paper presented at the annual meeting of AERA, Los Angeles, 1981.

[2] Draft copy. Not for circulation without the express permission of the authors.

DRAFT COPY. Not for circulation without the express permission of the authors

IRT Versus Conventional Equating Methods: A Comparative Study of Scale Stability

Nancy S. Petersen, Linda L. Cook, and Martha L. Stocking

Educational Testing Service, Princeton, New Jersey

Most major testing programs administer multiple forms of the same test. The various forms contain different questions but are carefully constructed to be as similar in difficulty and content as possible. It is inevitable, however, that there will exist some form-to-form differences. Therefore, if some assessment of relative ability is to be made for examinees who have taken different forms of the same test, it is necessary, for the sake of equity, to define a process that will render scores on these forms comparable. This process, which is referred to as equating, yields a transformation of raw scores to scaled scores (scores on an arbitrarily chosen common scale). Ideally, the end result of the equating process is that an examinee would receive the same scaled score regardless of which form of the test he or she was administered.

All equating methods have two components, a design for data collection and a statistical model for analyzing data. An external anchor-test design (Angoff 1971) was used in this study for data collection. An anchor-test design requires administering one form of a total test to one group of examinees, a second form to a second group of examinees, and a common anchor test to both groups. The anchor test may be included within the total test (internal anchor) or it may be administered separately (external anchor). If scores on two test forms are to be equated through their relations to scores on an anchor test, confidence in the results is

enhanced when the forms are parallel in all respects, when the anchor test is a miniature but otherwise parallel version of the total test, and when the equating operations are based on randomly equivalent samples.

In actual testing practice, however, these conditions cannot always be completely satisfied. Program policy for some nationally administered tests prohibits administering different test forms to random samples of examinees on the same test date. The composition of successive groups of examinees who elect to take a particular test tends to shift over time. There are usually differences even among the groups of examinees who are tested on different national administration dates within a single year, differences that occur for generally unknown but undoubtedly complex reasons. Furthermore, the specifications that guide test construction evolve over time, so that test forms used during different periods of a testing program's history may differ in important respects. Even when two test forms are themselves parallel, various practical constraints may require that their scores be equated through scores on an anchor test that does not faithfully reflect the characteristics of the total test.

Because test scores usually must be equated under the less-than-optimum conditions described above, this study sought to compare the extent to which different equating methods lead to scale drift. Tests are equated in order to maintain comparability of scores over time. Scale drift will occur if the result of equating, say, Form A (the new form) to Form D (the old form) directly is not the same as that obtained by equating Form A to Form D through intervening Forms B and C. More specifically, the purpose of this study was to compare the extent to which different equating methods lead to scale drift for the verbal and mathematical sections of the Scholastic Aptitude Test (SAT).

The Scholastic Aptitude Test is administered by the College Board Admissions Testing Program. The SAT is a multiple choice test that measures developed verbal and mathematical reasoning ability related to successful college performance. It is used extensively, in conjunction with measures of high school performance, to select students for admission to colleges.

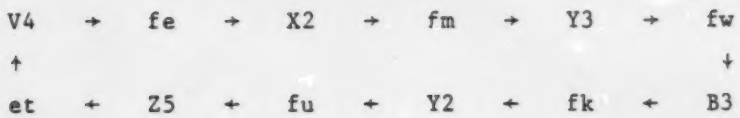
Scale drift for the verbal and mathematical sections of the SAT was investigated using linear, equipercentile, and item response theory (IRT) equating methods. Three linear models, namely, Tucker, Levine Equally Reliable, and Levine Unequally Reliable, were used. Three IRT calibration designs were investigated. The IRT designs will be referred to as (1) concurrent, (2) partial pre-calibration with fixed b's, and (3) partial pre-calibration with equated b's. Figure 1 illustrates the chain of six verbal equatings and the chain of six mathematical equatings that were used for the study. Upper case letters indicate SAT forms and lower case letters indicate anchor tests. It is possible to use each chain to assess scale drift in the following manner. Form V4, which has previously been put on scale, can be equated to itself through the five intervening forms X2, Y3, B3, Y2, and Z5. Any discrepancy between the transformation obtained from the circular chain of equatings and the initial V4 scale (the criterion scale) could be attributed to scale drift.

Study Design

Description of Tests

Test specifications for the SAT have changed over the years. Forms, such as V4, administered prior to the Fall of 1974 consist of two 45-minute sections (one verbal and one mathematical) and three 30-minute sections (one verbal, one mathematical, and one experimental containing an anchor

Chain of Six Verbal Equatings



Chain of Six Mathematical Equatings

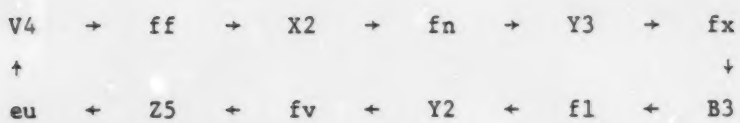


Figure 1: Verbal and Mathematical equating chains. Upper-case letters designate SAT forms and lower-case letters, anchor tests.

test or pretest). The two verbal sections contain a total of 90 5-choice items composed of 18 antonyms, 19 analogies, 18 sentence completions, and 7 reading passages each of which is followed by 5 items based on the passage. The two mathematical sections contain a total of 60 5-choice items composed of 42 regular mathematics items and 18 data sufficiency items.

Forms, such as X2, Y3, B3, Y2, and Z5, administered since the Fall of 1974 consist of six 30-minute sections: two verbal sections, two mathematical sections, one Test of Standard Written English, and one experimental section. The two verbal sections contain a total of 85 5-choice items composed of 25 antonyms, 20 analogies, 15 sentence completions, and 5 reading passages each of which is followed by 5 items based on the passage. The two mathematical sections contain a total of 60 items[1] composed of 40 5-choice regular mathematics items and 20 4-choice quantitative comparisons items.

Content specifications for the anchor tests associated with the SAT forms used in the study have not changed over time. The verbal anchor tests each contain 40 items (10 items of each type). The mathematical anchor tests each contain 25 regular mathematics items.[2]

Raw scores obtained on the SAT are typically transformed to scaled scores on the College Board 200 to 800 scale via linear equating methods. SAT raw scores are formula scores that have been corrected for guessing. Raw scores are computed by the formula R-W/k, where R is the number of correct responses, W is the number of incorrect responses, and (k+1) equals the number of choices per item.

---

[1] Scores on SAT-M form Y3 are based on only 59 items due to a printing error in one item.

[2] Scores on the mathematical anchor test fn are based on only 24 items due to a printing error in one item.

Data Collection

A sample of approximately 2670 cases was selected for each pairing of an SAT form and an anchor test shown in Figure 1. The samples were randomly selected from data obtained at the respective test administrations. The test administration dates and summary statistics for each verbal and mathematical sample are given in Table 1.

Equating Design

In order to assess the magnitude of scale drift associated with an equating method, each test form in a chain (see Figure 2) was equated to the preceeding form. For example, the verbal form V4 was treated as the base form (or initial form) of the test. The first step then in the chain of verbal equatings was to equate form X2 to form V4 using anchor test fe. Next, Y3 was equated to X2 using anchor test fm. In the last step, V4 was equated to Z5 using anchor test et.

For each equating method used in the study, the transformation obtained from equating V4 to V4 through the five intervening forms was compared to the initial V4 scale. Any discrepancy between the transformation obtained from the circular chain of equatings and the initial V4 scale was considered to be scale drift attributable to the equating method.

Equating Methodology

The three equating methods used in the study for determining the scale transformations can be described as follows:

1. Equipercentile equating. Choose a transformation such that scores from the two tests will be equated if they correspond to the same percentile rank in some group of examinees.

Table 1

Raw Score Summary Statistics for Verbal and Mathematical Equating Samples

Form[a]	Admin Date	N	SAT Form		Anchor Test		Covariance
			Mean	SD	Mean	SD	
Verbal Samples							
V4-fe	12/73	2665	35.04	16.37	14.01	8.54	123.22
X2-fe	4/75	2686	35.24	15.27	13.65	7.95	104.84
X2-fm	4/75	2562	34.42	15.31	16.74	8.07	106.75
Y3-fm	6/76	2578	34.48	16.34	16.14	8.41	121.17
Y3-fw	1/78	2549	31.37	15.86	14.36	8.17	114.28
B3-fw	5/79	2700	36.40	15.80	16.38	8.06	111.58
B3-fk	5/79	2665	35.90	15.24	15.04	8.01	106.07
Y2-fk	4/76	2879	34.16	14.84	15.08	8.19	105.23
Y2-fu	4/76	2774	33.57	14.50	16.02	7.44	92.29
Z5-fu	12/77	2853	30.73	15.61	14.43	7.69	103.94
Z5-et	12/77	2814	31.13	15.91	13.76	7.83	108.49
V4-et	12/73	2670	34.66	16.11	15.04	7.94	110.41
Mathematical Samples							
V4-ff	12/73	2628	20.74	12.89	9.46	5.55	59.91
X2-ff	4/75	2629	24.42	12.88	9.97	5.45	58.97
X2-fn	4/75	2527	23.84	12.90	9.73	5.73	63.24
Y3-fn	6/76	2553	24.05	13.30	9.57	5.85	67.40
Y3-fx	1/78	2455	21.48	13.74	8.47	6.33	77.11
B3-fx	5/79	2633	24.70	13.33	10.14	6.10	70.09
B3-f1	5/79	2596	24.30	12.97	9.46	5.36	59.09
Y2-f1	4/76	2815	23.25	13.49	9.13	5.46	62.83
Y2-fv	4/76	2721	23.68	13.46	9.76	6.01	69.98
Z5-fv	12/77	2774	20.52	13.49	8.38	5.96	69.73
Z5-eu	12/77	2739	20.13	13.38	8.53	5.92	68.77
V4-eu	12/73	2673	20.55	12.96	9.46	6.06	67.61

[a]Upper-case letters designate SAT forms and lower-case letters, anchor tests.

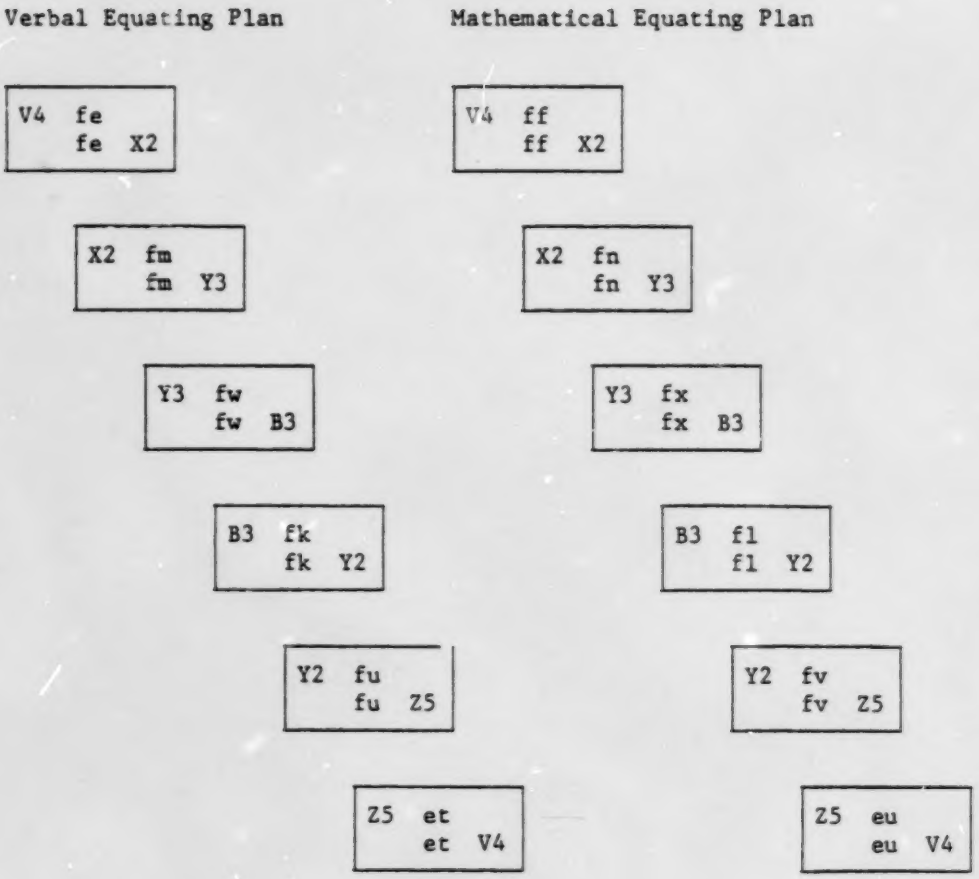


Figure 2: Verbal and Mathematical equating plans. Boxes indicate separate equating runs. Each pairing of an SAT form and an anchor test represents a sample of approximately 2670 cases.

2. Linear equating. Choose a linear transformation such that scores from the two tests will be equated if they correspond to the same number of standard deviations from the mean in some group of examinees.
3. Item response theory equating. Choose a transformation such that scores from the two tests will be equated if they correspond to the same estimated level of the latent trait, ability, or skill underlying both tests.

Equipercentile Equating

In equipercentile equating (Angoff, 1971, p. 571), scores on each test form and on the anchor test are first equated separately within each group. Then scores on the two tests to be equated are said to be equivalent if they correspond to the same score on the anchor test.

Linear Equating

Linear equating methods all produce an equating transformation of the form $T(x) = Ax + B$, where T is the equating transformation, x is the test score to which it is applied, and A and B are parameters estimated from the data. The Tucker, Levine Equally Reliable, and Levine Unequally Reliable linear equating models (Angoff, 1971, pp. 579-583) were used in this study. These models are based on univariate selection sampling theory. Scores on the relevant selection attribute (the attribute on which the equating samples vary) are assumed to be collinear with scores on the anchor test in the case of the Tucker model and with true scores on both the anchor test and the test form in the case of the Levine models. Scores on the anchor test are used to estimate performance of the combined group

of examinees on both the old and new forms of the test; thus, simulating by statistical methods, the situation in which the same group of examinees takes both forms of the test.

The parameters A and B of the equating transformation are estimated by means of an equation that expresses the idea of equating in standard score terms:

$$(x - m_x)/s_x = (y - m_y)/s_y \quad , \qquad (1)$$

where x and y refer to the test scores to be equated, and m and s refer to the means and standard deviations of the scores in some group of examinees. Methods using the above equation differ in their identification of the means and standard deviations to be estimated. The Tucker and Levine Equally Reliable methods are based on the estimated means and standard deviations of observed scores whereas the Levine Unequally Reliable method is based on the estimated means and standard deviations of true scores.

The formulas for computing the A and B parameters for the Tucker, Levine Equally Reliable, and Levine Unequally Reliable models are given in Table 2. As noted in Table 2, the formulas for the Levine models require error variance estimates. Angoff's method (1953) of estimating error variances was used in this study. This method assumes that the test to be equated and the anchor test are parallel except for length.

Item Response Theory Equating

Item response theory (IRT) assumes that there is a mathematical function which relates the probability of a correct response on an item to an examinee's ability. (See Lord, Chapter 2, 1980 for a detailed

Table 2

Formulas for Linear Conversion Parameters

Tucker

$$A = (S^2_{yb} + C^2_{yvb}(S^2_{vc} - S^2_{vb})/S^4_{vb})^{\frac{1}{2}}(S^2_{xa} + C^2_{xva}(S^2_{vc} - S^2_{va})/S^4_{va})^{-\frac{1}{2}}$$

$$B = M_{yb} + C_{yvb}(M_{vc} - M_{vb})/S^2_{vb} - AM_{xa} - AC_{xva}(M_{vc} - M_{va})/S^2_{va}$$

Levine Equally Reliable

$$A = (S^2_{yb} + (S^2_{yb} - S^2_{y''b})(S^2_{vc} - S^2_{vb})/(S^2_{vb} - S^2_{v''b}))^{\frac{1}{2}}$$
$$(S^2_{va} + (S^2_{xa} - S^2_{x''a})(S^2_{vc} - S^2_{va})/(S^2_{va} - S^2_{v''a}))^{-\frac{1}{2}}$$

$$B = M_{yb} + (M_{vc} - M_{vb})((S^2_{yb} - S^2_{y''b})/(S^2_{vb} - S^2_{v''b}))^{\frac{1}{2}}$$
$$- AM_{xa} - A(M_{vc} - M_{va})((S^2_{xa} - S^2_{x''a})/(S^2_{va} - S^2_{v''a}))^{\frac{1}{2}}$$

Levine Unequally Reliable

$$A = ((S^2_{yb} - S^2_{y''b})/(S^2_{vb} - S^2_{v''b}))^{\frac{1}{2}}((S^2_{xa} - S^2_{x''a})/(S^2_{va} - S^2_{v''a}))^{-\frac{1}{2}}$$

$$B = M_{yb} + (M_{va} - M_{vb})((S^2_{yb} - S^2_{y''b})/(S^2_{vb} - S^2_{v''b}))^{\frac{1}{2}} - AM_{xa}$$

Angoff Error Variance Estimates (Anchor Test External to Total Test)

$$S^2_{p''g} = (S^2_{pg}S^2_{vg} - C^2_{pvg})/(S^2_{vg} + C_{pvg})$$

$$S^2_{v''g} = (S^2_{pg}S^2_{vg} - C^2_{pvg})/(S^2_{pg} + C_{pvg})$$

Notation

New Test Form	X
Old Test Form	Y
Either New or Old Test Form	P
Anchor Test	V
Observed Score	x, y, v, p
Error Score	x", y", v", p"
Group Taking Test X and Test V	a
Group Taking Test Y and Test V	b
Group Taking Test P and Test V	g
Combined Group	c or (a + b)
Mean	M
Standard Deviation	S
Covariance	C

discussion). Many different mathematical models of this functional relationship are possible. The model chosen for this study was the three-parameter logistic model. In this model, where θ represents an examinee's ability, the probability of a correct response to item i, $P_i(\theta)$, is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.702a_i(\theta - b_i)}} \quad ,$$

where a_i, b_i, and c_i are three parameters describing the item. These parameters have specific interpretations: b_i is the point on the θ metric at the inflection point of $P_i(\theta)$ and is interpreted as the item difficulty; a_i is proportional to the slope of $P_i(\theta)$ at the point of inflection and represents the item discrimination; and c_i is the lower asymptote of $P_i(\theta)$ and represents a pseudo-guessing parameter.

The item parameters and examinee abilities for this study were estimated (calibrated) using the program LOGIST (Wood and Lord, 1976; Wood et al., 1976). The estimates are obtained by a (modified) maximum likelihood procedure with special procedures for the treatment of omitted items (see Lord, 1974).

LOGIST requires as input the responses to a set of items from a group of examinees, coded to reflect items answered correctly, incorrectly, omitted, and not reached. In addition, the user may specify certain restrictions on the data and parameters in order to speed convergence of the iterative procedure. The major restrictions specified for the study for most LOGIST computer runs were:

1. examinees who answered less than one-third of the items were not used,
2. a's were restricted to a range of .01 to 1.5,
3. c's were restricted to a range of .09 to .20, and
4. θ's were restricted to a range of -7.0 to 4.0.

LOGIST produces as output estimates of the a, b, and c for each item, and θ for each examinee. The metric chosen arbitrarily for the θ (and b) scale is such that the distribution of estimates of θ has mean zero and standard deviation one. If two separate LOGIST runs are made for the same items, but different groups of examinees, the resulting parameter estimates will be on different scales. There will be, however, a simple linear relationship that transforms one scale to the other.

Three different item calibration designs were studied. They will be referred to as (a) concurrent calibration, (b) partial pre-calibration with fixed b's, and (c) partial pre-calibration with equated b's.

For the concurrent calibration design, each pair of verbal or mathematical test forms plus associated anchor test is calibrated in a single LOGIST run (see Figure 2). This results in item parameters on a common scale for each pair and allows direct equating of the two forms.

For the two partial pre-calibration designs, LOGIST runs were made separately for each test form and associated anchor test (see Figure 3). Two different procedures were used to place all item parameters on the same scale.

In one procedure (fixed b's), items were calibrated sequentially, starting with the V4-fe box in the verbal chain and with the V4-ff box in the mathematical chain in Figure 3. For the second LOGIST run in the verbal chain, the item difficulties for the fe section obtained from the

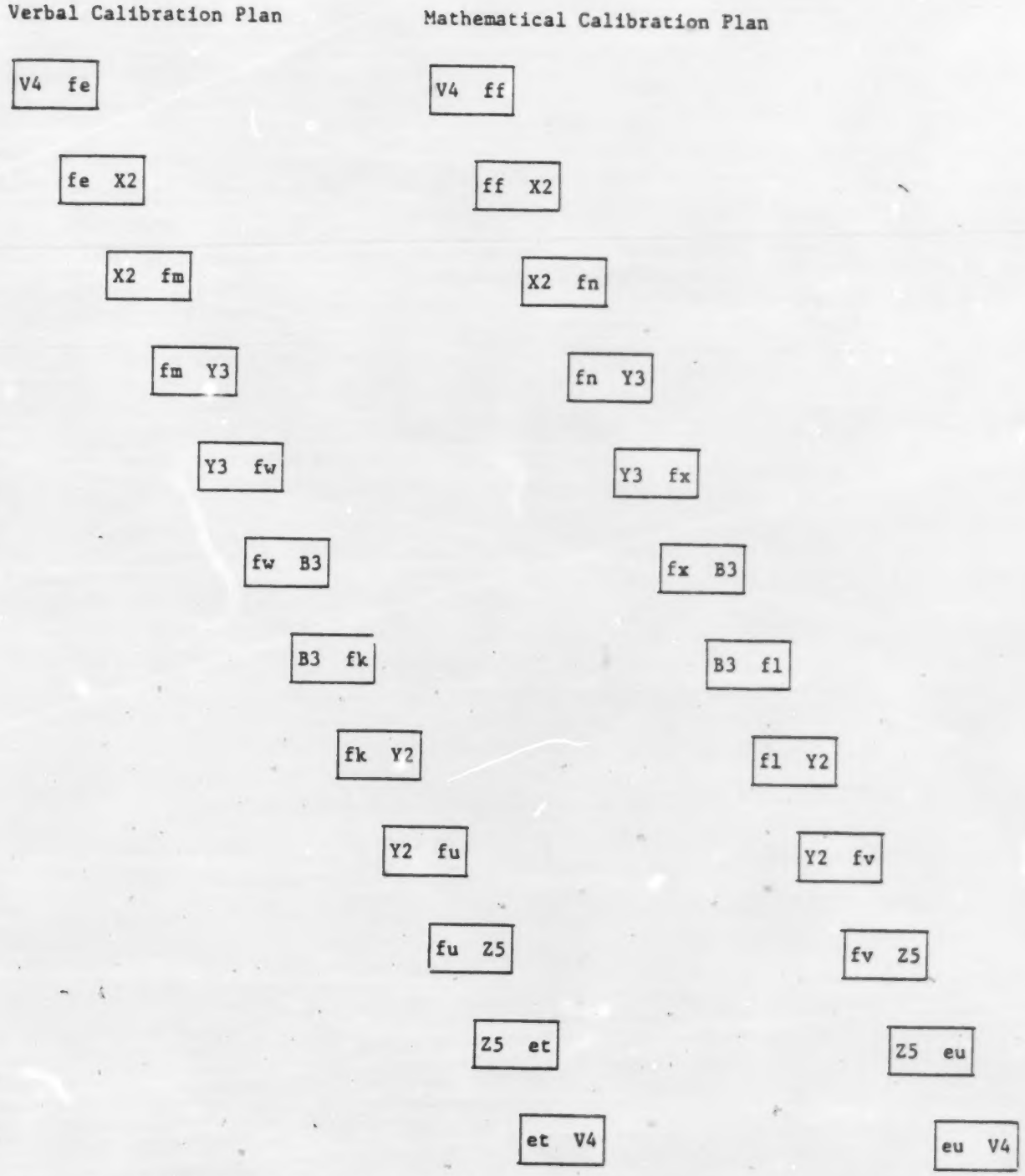


Figure 3: Verbal and Mathematical partial pre-calibration plans. Boxes indicate separate calibration runs. Each pairing of an SAT form and an anchor test represents a sample of approximately 2670 cases.

first run were used as input and were not re-estimated. The parameter estimates for X2 are then automatically placed on the same scale as those for fe. This procedure was continued down each chain with b's from items common with the previous LOGIST run being held fixed, placing the new items for the current run on the same scale. When each chain is complete, all parameter estimates within a chain are now on the same scale, and equating of scores on each pair of test forms can be performed.

In the other partial pre-calibration procedure (equated b's), each pairing of an SAT form and an anchor test shown in Figure 3 was calibrated in an independent LOGIST run. To place all estimates of item parameters within a chain on a common scale, a sequential transformation process was employed (see Appendix A). For example, the first two LOGIST runs for the verbal chain produced estimates for V4-fe on one scale, and for fe-X2 on a different scale. The two separate item difficulty estimates for items in section fe were then related by mean and sigma equating using weighted estimates of the mean and standard deviation. The weights are a combination of the standard error of the estimated item difficulties and weights designed to decrease the influence of outliers. This procedure yields the transformation that needs to be applied to the a and b parameters for all items from the fe-X2 calibration to place them on the same scale as V4-fe. The procedure is repeated, using the transformation obtained from the relationship between the difficulty estimates for form X2 to place X2-fm items on the same scale. The transformations were continued sequentially down each chain resulting in a common scale for all item parameter estimates within a chain. Equating of each pair of test forms can then be performed.

Although there are a number of equating techniques possible when using IRT, this study was concerned only with true formula score equating (Lord, Chapter 13, 1980). The expected value of an examinee's observed formula score is defined as his or her true formula score. For the true formula score, ξ, we have

$$\xi = \sum_{i=1}^{n} \left[\frac{(k_i+1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right]$$

where n is the number of items in the test and (K_i+1) is the number of choices for item i. If we have two tests measuring the same ability θ, then true formula scores ξ and η from the two tests are related by the equations

$$\begin{aligned} \xi &= \sum_{i=1}^{n} \left[\frac{(k_i+1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right] \\ \eta &= \sum_{j=1}^{m} \left[\frac{(k_j+1)}{k_j} P_j(\theta) - \frac{1}{k_j} \right] \end{aligned} \qquad (2)$$

Clearly, for a particular θ corresponding true scores ξ and η have identical meaning. They are said to be equated.

Because true formula scores below the chance score level are undefined for the three-parameter logistic model, some method must be established to obtain a relationship between scores below the chance level on the two test forms to be equated. The approach used for this study (Lord, 1980) was to estimate the mean (m) and standard deviation (s) of below chance level scores on the two tests to be equated via the following formulas:

$$m = \sum_{i=1}^{n} (c_i(k_i+1)/k_i - 1/k_i) \qquad (3)$$

$$s^2 = \sum_{i=1}^{n} (c_i - c_i^2)\,(k_i+1)^2/k_i^2$$

where n is the number of items in the test, (k_i+1) is the number of choices for item i, and c_i is the psuedo-guessing parameter for item i; and then to use these estimates to do a simple linear equating (see equation (1)) between the two sets of below chance level scores.

In practice, true score equating is carried out by substituting estimated parameters into the equations (2) and (3). Paired values of ξ and η are then computed for a series of arbitrary values of θ. Since we cannot know an examinee's true formula score, we act as if relationships (2) and (3) apply to an examinee's observed formula score.

For the concurrent calibration design (see Figure 2), the equating procedure was applied sequentially starting with the items calibrated in the first LOGIST run for each chain. Parameters were already available to convert raw scores on verbal and mathematical forms V4 to the College Board scale. Equivalent true formula score estimates were found for V4 and X2, resulting in a table of transformations of raw scores on X2 to the College Board scale. Form Y3 was then (true formula score) equated to X2, resulting in a table of transformations for raw scores on Y3 to the College Board scale. This procedure was repeated sequentially down the verbal and mathematical chains. The end product is a table of transformations of the raw scores on form V4 to the College Board scale.

For the partial pre-calibration designs (see Figure 3), a sequential equating procedure is not necessary since all item parameter estimates are on the same scale. The intermediate equatings can be ignored and

only the equating of form V4 to form V4 need be performed. This is not to say that comparisons of these intermediate equatings between the two partial pre-calibration designs might not give some insight into the effects of the two scaling procedures.

The final and initial conversions of V4 to the College Board scale should, of course, be identical for all equating procedures. Departures resulting in scale drift may be due to sampling error or model fit problems.

Results

The initial and final raw score to scale score transformations for form V4 are listed in Table 3 for SAT-Verbal and in Table 4 for SAT-Mathematical. Also given in Tables 3 and 4 are the raw score frequencies for the total group taking form V4 in December 1973. The equipercentile conversions do not extend over the possible raw score range because equipercentile equating can only be done over that score range for which there are observations in the equating samples. Theoretically, the initial and final conversion of V4 raw scores to the College Board scale should be identical for each equating method. To illustrate the extent to which the initial (criterion) and final conversions differ, the raw score to scale score conversions listed in Tables 3 and 4 are shown graphically in Figures 4 and 5, respectively.

Perusal of the SAT-Verbal conversions shown in Figure 4 indicates that the final linear, equipercentile, IRT concurrent, and IRT fixed b's conversions tend to overestimate the initial scale values whereas the IRT equated b's conversion tends to underestimate the initial scale values. The linear conversions are more similar to each other than they are to the IRT conversions. Of the IRT conversions, the concurrent and fixed b's are the most similar. The equated b's conversion is the most similar

TABLE 3

INITIAL AND FINAL TRANSFORMATIONS
OF SAT-VERBAL FORM V4 RAW SCORES TO SCALED SCORES

RAW SCORE	A FREQ	INITIAL SCALED SCORE	ESTIMATED SCALED SCORE						
			TUCKER	LEVINE EQ REL	LEVINE UNEQ REL	EQUITILE	CONCURRENT	FIXED B	EQUATED B
90	3	822.62	856.64	850.77	853.53		822.61	822.61	822.61
89	9	815.63	849.41	843.61	846.32		821.64	816.25	815.77
88	11	808.63	842.18	836.44	839.11		818.56	810.25	809.05
87	1	801.64	834.94	829.28	831.90		814.25	804.36	802.43
86	26	794.65	827.71	822.11	824.68		809.20	798.52	795.87
85	44	787.66	820.47	814.95	817.47		803.69	792.69	789.35
84	71	780.66	813.24	807.78	810.26		797.89	786.83	782.84
83	98	773.67	806.01	800.62	803.05		791.90	780.95	776.34
82	38	766.68	798.77	793.46	795.84		785.78	775.03	769.82
81	135	759.68	791.54	786.29	788.63		779.55	769.05	763.27
80	172	752.69	784.30	779.13	781.42	772.34	773.20	763.02	756.70
79	215	745.70	777.07	771.96	774.21	763.81	766.75	756.93	750.09
78	251	738.70	769.84	764.80	767.00	754.02	760.18	750.78	743.45
77	164	731.71	762.60	757.63	759.79	744.93	753.51	744.56	736.77
76	294	724.72	755.37	750.47	752.58	738.98	746.75	738.28	730.05
75	360	717.72	748.13	743.30	745.36	724.65	739.88	731.94	723.28
74	441	710.73	740.90	736.14	738.15	718.63	732.92	725.54	716.47
73	533	703.74	733.67	728.98	730.94	715.56	725.87	719.07	709.62
72	374	696.75	726.43	721.81	723.73	712.83	718.75	712.54	702.72
71	590	689.75	719.20	714.65	716.52	706.19	711.54	705.94	695.77
70	667	682.76	711.96	707.48	709.31	696.40	704.20	699.27	688.77
69	791	675.77	704.73	700.32	702.10	693.32	696.90	692.55	681.73
68	913	668.77	697.50	693.15	694.89	690.87	689.63	685.75	674.65
67	650	661.78	690.26	685.99	687.68	688.63	682.24	678.89	667.52
66	1002	654.79	683.03	678.83	680.47	686.40	674.81	671.97	660.34
65	1166	647.79	675.79	671.66	673.26	684.37	667.34	664.99	653.13
64	1313	640.80	668.56	664.50	666.04	681.92	659.87	657.94	645.88
63	1486	633.81	661.33	657.33	658.83	677.93	652.37	650.84	638.60
62	1022	626.81	654.09	650.17	651.62	665.21	644.86	643.68	631.29
61	1562	619.82	646.86	643.00	644.41	646.12	637.34	636.47	623.95
60	1801	612.83	639.62	635.84	637.20	636.46	629.81	629.21	616.59
59	1955	605.84	632.39	628.68	629.99	627.93	622.26	621.91	609.22
58	2182	598.84	625.16	621.51	622.78	622.69	614.73	614.56	601.83
57	1633	591.85	617.92	614.35	615.57	617.93	607.19	607.18	594.41
56	2228	584.86	610.69	607.18	608.36	611.22	599.68	599.78	587.02
55	2583	577.86	603.45	600.02	601.15	604.72	592.20	592.35	579.61
54	2694	570.87	596.22	592.85	593.94	597.30	584.70	584.89	572.19
53	2987	563.88	588.99	585.69	586.73	587.02	577.24	577.43	564.79
52	2202	556.88	581.75	578.53	579.51	578.07	569.80	569.95	557.38
51	3133	549.89	574.52	571.36	572.30	572.06	562.39	562.47	549.99
50	3474	542.90	567.28	564.20	565.09	566.39	555.01	555.00	542.62
49	3653	535.90	560.05	557.03	557.88	561.08	547.67	547.53	535.25
48	3998	528.91	552.82	549.87	550.67	555.48	540.36	540.07	527.91
47	3011	521.92	545.58	542.70	543.46	549.82	533.09	532.63	520.59
46	4194	514.92	538.35	535.54	536.25	545.06	525.80	525.22	513.29
45	4418	507.93	531.11	528.38	529.04	541.57	518.73	517.83	506.03
44	4724	500.94	523.88	521.21	521.83	538.07	511.59	510.46	498.78

TABLE 3 (CONTINUED)

INITIAL AND FINAL TRANSFORMATIONS
OF SAT-VERBAL FORM V4 RAW SCORES TO SCALED SCORES

RAW SCORE	A FREQ	INITIAL SCALED SCORE	ESTIMATED SCALED SCORE						
			TUCKER	LEVINE EQ REL	LEVINE UNEQ REL	EQUIPERCENTILE	CONCURRENT	FIXED B	EQUATED B
43	4448	493.95	516.65	514.05	514.62	531.08	504.49	503.12	491.56
42	3793	486.95	509.41	506.88	507.41	523.60	497.44	495.82	484.37
41	5032	479.96	502.18	499.72	500.19	515.27	490.42	488.55	477.22
40	5449	472.97	494.94	492.55	492.98	504.30	483.44	481.32	470.09
39	5492	465.97	487.71	485.39	485.77	492.69	476.48	474.12	462.99
38	5435	458.98	480.48	478.23	478.56	479.96	469.56	466.96	455.91
37	4420	451.99	473.24	471.06	471.35	470.94	462.64	459.83	448.87
36	5794	444.99	466.01	463.90	464.14	463.25	455.74	452.73	441.84
35	5942	438.00	458.77	456.73	456.93	455.90	448.85	445.66	434.84
34	5933	431.01	451.54	449.57	449.72	445.48	441.95	438.62	427.86
33	6072	424.01	444.31	442.40	442.51	438.35	435.04	431.61	420.90
32	4537	417.02	437.07	435.24	435.30	431.36	428.11	424.61	413.95
31	5860	410.03	429.84	428.08	428.09	423.32	421.16	417.64	407.02
30	5996	403.04	422.60	420.91	420.87	414.64	414.14	410.68	400.10
29	6005	396.04	415.37	413.75	413.66	405.20	407.14	403.74	393.20
28	5990	389.05	408.14	406.58	406.45	396.60	400.05	396.81	386.30
27	4521	382.06	400.90	399.42	399.24	387.58	392.93	389.88	379.42
26	5359	375.06	393.67	392.25	392.03	379.61	385.74	382.96	372.54
25	5381	368.07	386.43	385.09	384.82	374.43	378.54	376.04	365.68
24	5325	361.08	379.20	377.93	377.61	370.10	371.25	369.11	358.81
23	5002	354.08	371.97	370.76	370.40	366.18	363.92	362.18	351.96
22	3719	347.09	364.73	363.60	363.19	361.08	356.55	355.26	345.12
21	4666	340.10	357.50	356.43	355.98	354.22	349.10	348.32	338.28
20	4491	333.10	350.26	349.27	348.77	346.81	341.60	341.38	331.46
19	4356	326.11	343.03	342.10	341.55	340.52	334.05	334.43	324.64
18	4086	319.12	335.80	334.94	334.34	334.99	326.44	327.46	317.83
17	2933	312.12	328.56	327.78	327.13	329.82	318.77	320.49	311.02
16	3472	305.13	321.33	320.61	319.92	323.10	311.05	313.49	304.22
15	3320	298.14	314.09	313.45	312.71	316.88	303.29	306.48	297.43
14	3112	291.15	306.86	306.28	305.50	308.42	295.48	299.45	290.64
13	2836	284.15	299.63	299.12	298.29	298.77	287.65	292.39	283.85
12	1967	277.16	292.39	291.95	291.08	290.73	279.80	285.30	277.06
11	2315	270.17	285.16	284.79	283.87	281.70	271.98	278.18	270.28
10	2209	263.17	277.92	277.63	276.66	271.42	264.18	271.02	263.50
9	1967	256.18	270.69	270.46	269.45	264.22	256.45	263.81	256.72
8	1754	249.19	263.46	263.30	262.23	258.84	248.78	256.55	249.94
7	1120	242.19	256.22	256.13	255.02	253.59	241.23	249.24	243.16
6	1290	235.20	248.99	248.97	247.81	233.52	233.81	241.88	236.39
5	1194	228.21	241.75	241.80	240.60	225.67	226.54	234.44	229.61
4	1057	221.21	234.52	234.64	233.39	220.72	219.44	226.93	222.82
3	853	214.22	227.29	227.48	226.18	202.96	212.52	219.33	216.03
2	484	207.23	220.05	220.31	218.97	193.38	205.78	211.64	209.23
1	522	200.24	212.82	213.15	211.76	191.70	199.22	203.85	202.41
0	484	193.24	205.58	205.98	204.55		192.84	195.94	195.56
-1	353	186.25	198.35	198.82	197.34		186.63	187.91	188.68
-2	245	179.26	191.12	191.65	190.13		180.59	179.75	181.78
-3	112	172.26	183.88	184.49	182.91		174.70	171.44	174.84

TABLE 3 (CONTINUED)

INITIAL AND FINAL TRANSFORMATIONS
OF SAT-VERBAL FORM V4 RAW SCORES TO SCALED SCORES

RAW SCORE	A FREQ	INITIAL SCALED SCORE	ESTIMATED SCALED SCORE						
			TUCKER	LEVINE EQ REL	LEVINE UNEQ REL	EQUIPERCENTILE	CONCURRENT	FIXED B	EQUATED B
-4	129	165.27	176.65	177.33	175.70		168.97	162.97	167.87
-5	91	158.28	169.41	170.16	168.49		163.67	154.26	160.90
-6	52	151.28	162.18	163.00	161.28		156.46	146.75	153.89
-7	10	144.29	154.95	155.83	154.07		149.24	139.89	146.90
-8	6	137.30	147.71	148.67	146.86		142.01	133.03	139.71
-9	14	130.30	140.48	141.50	139.65		134.79	126.17	132.62
-10	6	123.31	133.24	134.34	132.44		127.56	119.32	125.54
-11	3	116.32	126.01	127.18	125.23		120.34	112.46	118.45
-12	2	109.32	118.78	120.01	118.02		113.11	105.60	111.36
-13	0	102.33	111.54	112.85	110.81		105.99	98.74	104.27
-14	1	95.34	104.31	105.68	103.59		98.66	91.88	97.18
-15		88.35	97.07	98.52	96.38		91.44	85.02	90.09
-16		81.35	89.84	91.35	89.17		84.21	78.16	83.00
-17		74.36	82.61	84.19	81.96		76.99	71.30	75.92
-18		67.37	75.37	77.03	74.75		69.76	64.44	68.83
-19		60.37	68.14	69.86	67.54		62.54	57.58	61.74
-20		53.38	60.90	62.70	60.33		55.31	50.73	54.65
-21		46.39	53.67	55.53	53.12		48.09	43.87	47.56
-22		39.39	46.44	48.37	45.91		40.86	37.01	40.47

A
SAT-VERBAL FORM V4 RAW SCORE FREQUENCY DISTRIBUTION FOR INITIAL DECEMBER 1973 ADMINISTRATION.

TABLE 4

INITIAL AND FINAL TRANSFORMATIONS
OF SAT-MATH FORM V4 RAW SCORES TO SCALED SCORES

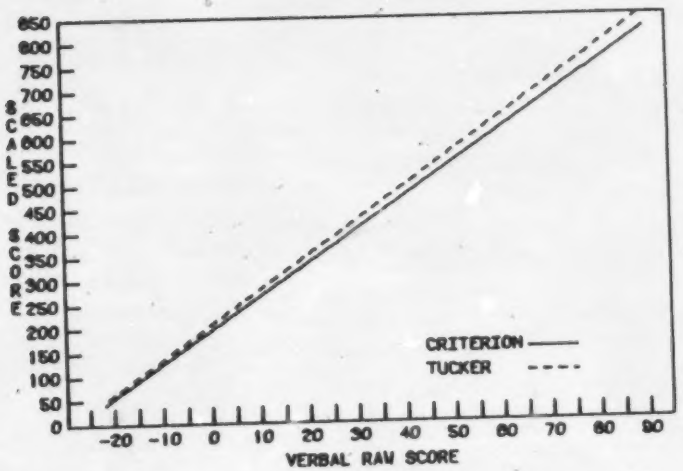
			ESTIMATED SCALED SCORE						
RAW SCORE	A FREQ	INITIAL SCALED SCORE	TUCKER	LEVINE EQ REL	LEVINE UNEQ REL	EQUIPERCENTILE	CONCURRENT	FIXED B	EQUATED B
60	34	805.63	817.43	812.74	815.32		805.62	805.62	805.62
59	114	796.74	808.47	803.78	806.30		799.33	798.79	796.40
58	205	787.86	799.50	794.82	797.28		792.03	791.76	786.76
57	47	778.97	790.54	785.87	788.24		784.29	784.64	777.15
56	317	770.09	781.57	776.91	779.24		776.37	777.43	767.66
55	475	761.21	772.61	767.94	770.22	759.96	768.37	770.11	758.29
54	638	752.32	763.64	759.00	761.23	756.23	760.32	762.67	749.00
53	795	743.44	754.68	750.05	752.18	753.30	752.23	755.09	739.78
52	430	734.56	745.72	741.09	743.16	747.35	744.03	747.35	730.59
51	878	725.67	736.75	732.14	734.14	731.45	735.73	739.44	721.43
50	1094	716.79	727.79	723.19	725.12	726.03	727.31	731.36	712.26
49	1312	707.90	718.82	714.22	716.10	717.05	718.77	723.12	703.09
48	1555	699.02	709.86	705.27	707.08	703.82	710.10	714.72	693.90
47	1144	690.14	700.89	696.31	698.06	693.33	701.32	706.15	684.70
46	1647	681.25	691.93	687.36	689.04	682.23	692.40	697.44	675.48
45	1939	672.37	682.96	678.40	680.02	669.17	683.37	688.59	666.23
44	2235	663.48	674.00	669.45	671.00	658.06	674.23	679.61	656.96
43	2452	654.60	665.03	660.49	661.98	648.83	664.98	670.52	647.67
42	2117	645.72	656.07	651.54	652.96	640.56	655.65	661.32	638.36
41	2588	636.83	647.10	642.58	643.94	633.13	646.23	652.02	629.01
40	2859	627.95	638.14	633.63	634.92	626.53	636.76	642.63	619.65
39	3213	619.06	629.18	624.67	625.90	619.06	627.24	633.17	610.26
38	3560	610.18	620.21	615.71	616.88	612.85	617.69	623.65	600.85
37	3059	601.30	611.25	606.76	607.86	606.54	608.13	614.08	591.43
36	3838	592.41	602.28	597.80	598.84	600.85	598.57	604.47	582.00
35	4022	583.53	593.32	588.85	589.82	595.79	589.02	594.84	572.57
34	4385	574.65	584.35	579.89	580.80	588.86	579.49	585.18	563.15
33	4761	565.76	575.39	570.94	571.78	577.93	570.01	575.53	553.74
32	3617	556.88	566.42	561.98	562.76	566.56	560.59	565.88	544.34
31	5053	547.99	557.46	553.03	553.74	553.41	551.23	556.25	534.97
30	5248	539.11	548.49	544.07	544.72	540.53	541.96	546.64	525.62
29	5511	530.23	539.53	535.11	535.70	529.07	532.77	537.07	516.30
28	5870	521.34	530.57	526.16	526.68	517.97	523.67	527.55	507.00
27	4779	512.46	521.60	517.20	517.66	508.28	514.67	518.06	497.73
26	6034	503.57	512.64	508.25	508.64	500.46	505.75	508.63	488.48
25	6466	494.69	503.67	499.29	499.62	494.60	496.92	499.24	479.25
24	6576	485.81	494.71	490.34	490.60	489.63	488.17	489.89	470.03
23	6894	476.92	485.74	481.38	481.58	484.30	479.47	480.58	460.82
22	5470	468.04	476.78	472.43	472.56	477.10	470.83	471.31	451.61
21	6934	459.15	467.81	463.47	463.54	469.46	462.22	462.08	442.40
20	6844	450.27	458.85	454.52	454.52	461.73	453.64	452.86	433.18
19	7185	441.39	449.88	445.56	445.50	454.36	445.06	443.67	423.96
18	7204	432.50	440.92	436.60	436.49	443.34	436.48	434.50	414.71
17	5444	423.62	431.95	427.65	427.46	429.57	427.88	425.33	405.45
16	6309	414.73	422.99	418.69	418.44	419.62	419.25	416.17	396.17
15	6986	405.85	414.03	409.74	409.47	411.54	413.58	407.01	386.87
14	7014	396.97	405.04	400.78	400.40	402.21	401.86	397.85	377.56

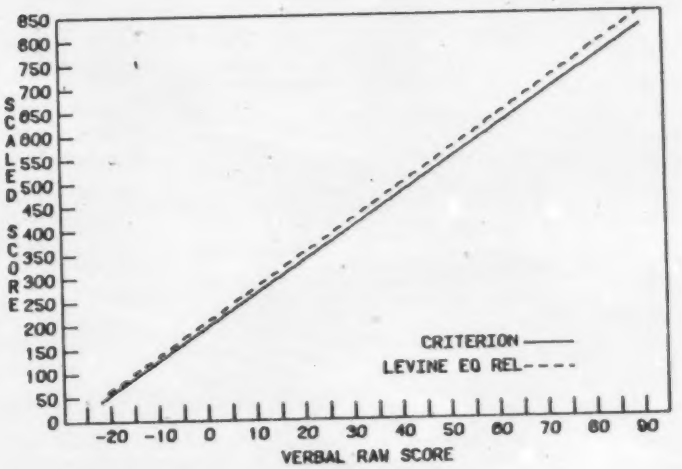
TABLE 4 (CONTINUED)

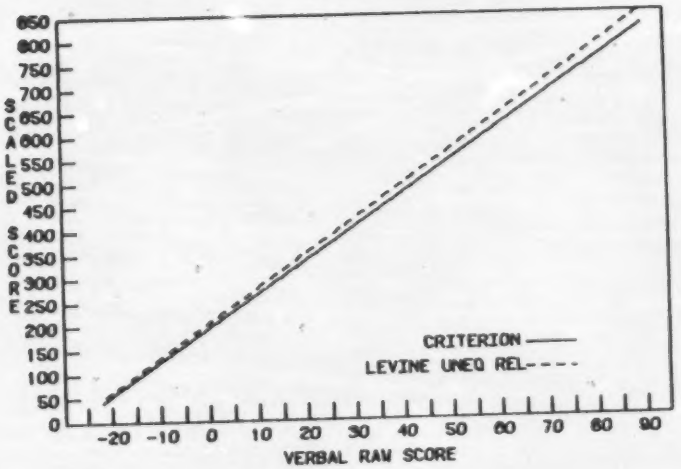
INITIAL AND FINAL TRANSFORMATIONS
OF SAT-MATH FORM V4 RAW SCORES TO SCALED SCORES

RAW SCORE	A FREQ	INITIAL SCALED SCORE	ESTIMATED SCALED SCORE						
			TUCKER	LEVINE EQ REL	LEVINE UNEQ REL	EQUIPERCENTILE	CONCURRENT	FIXED B	EQUATED B
13	6696	384.38	396.13	391.83	391.38	387.11	393.08	388.69	368.24
12	5006	379.20	387.13	382.87	382.36	372.00	384.24	379.52	358.90
11	6025	370.32	378.17	373.92	373.34	364.45	375.32	370.34	349.57
10	5949	361.43	369.20	364.96	364.32	357.26	366.31	361.16	340.25
9	5588	352.55	360.24	356.00	355.30	351.30	357.19	351.98	330.95
8	5411	343.66	351.27	347.05	346.28	344.46	347.94	342.80	321.69
7	3576	334.78	342.31	338.09	337.26	338.42	338.55	333.63	312.47
6	4577	325.90	333.34	329.14	328.24	331.85	328.97	324.47	303.33
5	4155	317.01	324.38	320.18	319.22	324.92	319.20	315.32	294.27
4	3849	308.13	315.41	311.23	310.20	318.08	309.19	306.20	285.32
3	3405	299.24	306.45	302.27	301.18	309.19	298.95	297.11	276.51
2	2065	290.36	297.49	293.32	292.16	301.38	288.49	288.07	267.84
1	2448	281.48	288.52	284.36	283.14	294.36	277.84	279.07	259.35
0	2049	272.59	279.56	275.41	274.12	286.18	267.10	270.14	251.06
-1	1421	263.71	270.59	266.45	265.10	279.88	256.39	261.28	242.97
-2	1152	254.82	261.63	257.49	256.08	276.50	245.88	252.50	235.10
-3	456	245.94	252.66	248.54	247.06	273.75	235.72	243.79	227.45
-4	500	237.06	243.70	239.58	238.04	272.06	225.97	235.15	220.01
-5	324	228.17	234.73	230.63	229.02	270.11	216.40	226.51	212.74
-6	186	219.29	225.77	221.67	220.00	258.73	206.91	217.67	205.57
-7	97	210.41	216.80	212.72	210.98	244.08	200.43	208.25	197.94
-8	17	201.52	207.84	203.76	201.96	225.42	192.07	199.48	189.64
-9	17	192.64	198.87	194.81	192.94		183.71	190.70	181.35
-10	4	183.75	189.91	185.85	183.92		175.34	181.93	173.06
-11		174.87	180.95	176.89	174.89		166.98	173.15	164.77
-12		165.99	171.98	167.94	165.87		158.61	164.37	156.48
-13		157.10	163.02	158.98	156.85		150.25	155.60	148.18
-14		148.22	154.05	150.03	147.83		141.89	146.82	139.89
-15		139.33	145.09	141.07	138.81		133.52	138.05	131.60
-16		130.45	136.12	132.12	129.79				
-17		121.57	127.16	123.16	120.77				

A
SAT-MATH FORM V4 RAW SCORE FREQUENCY DISTRIBUTION FOR INITIAL DECEMBER 1973 ADMINISTRATION.







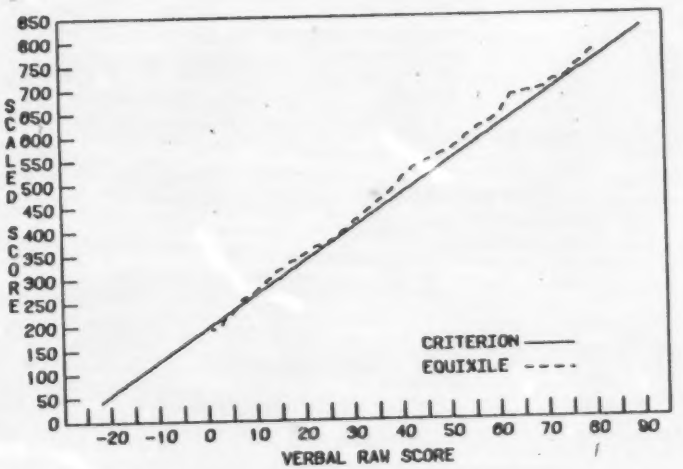
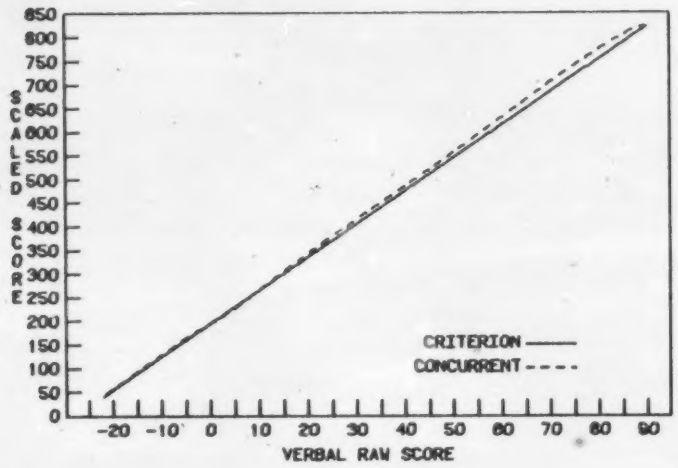
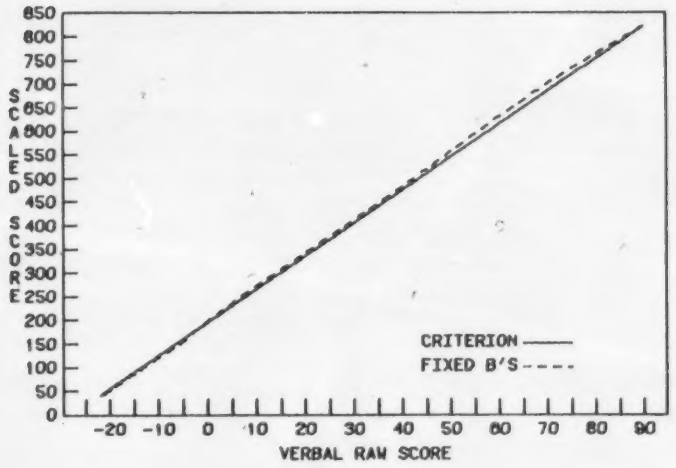


Figure 4: SAT-Verbal chain. Comparison of initial scale (criterion) and final conversion for each equating method.





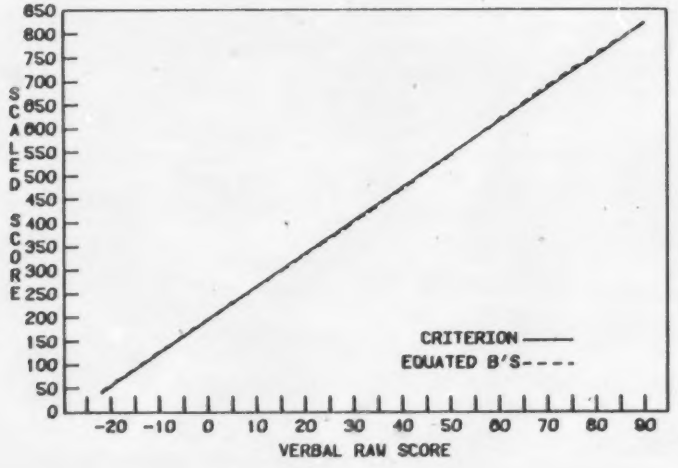
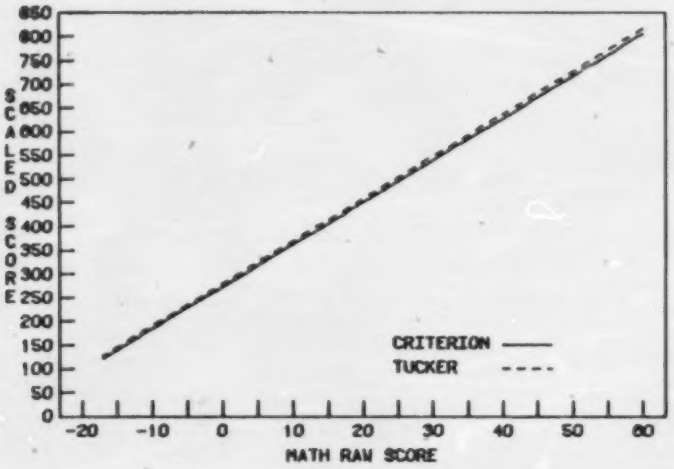
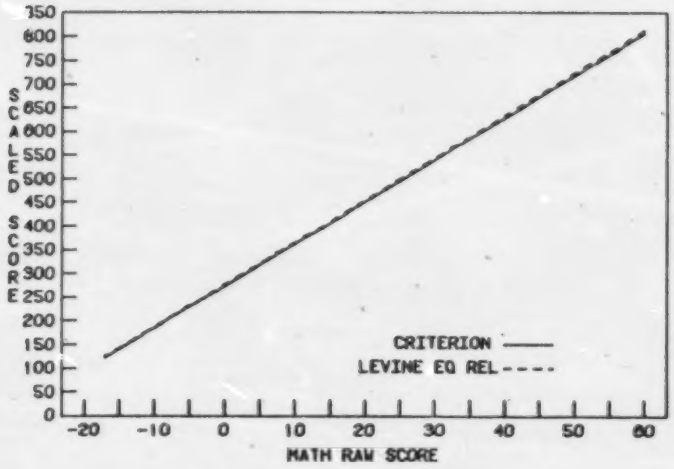
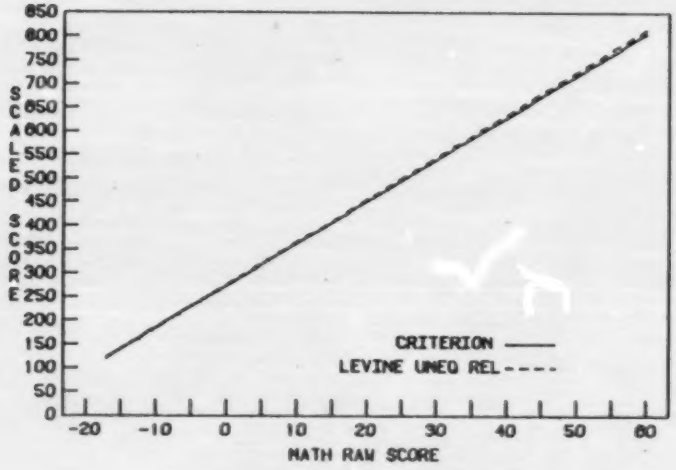


Figure 4 (continued)







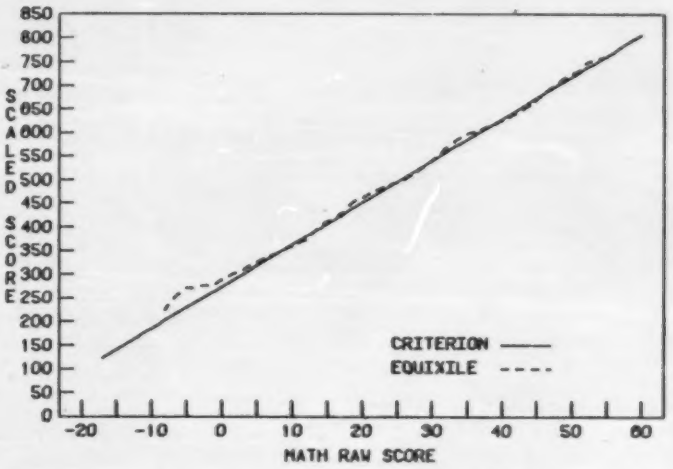
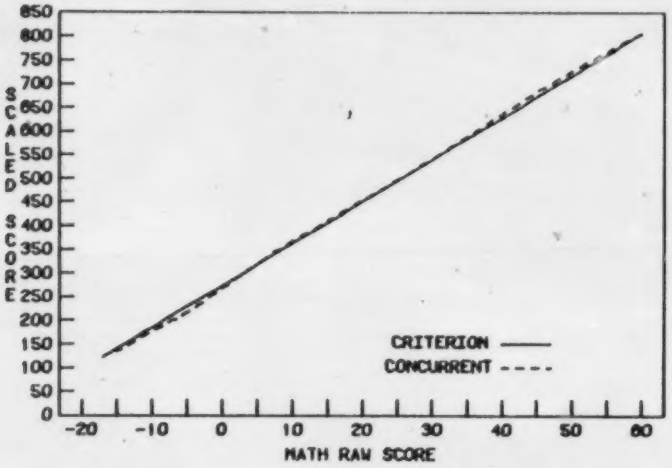
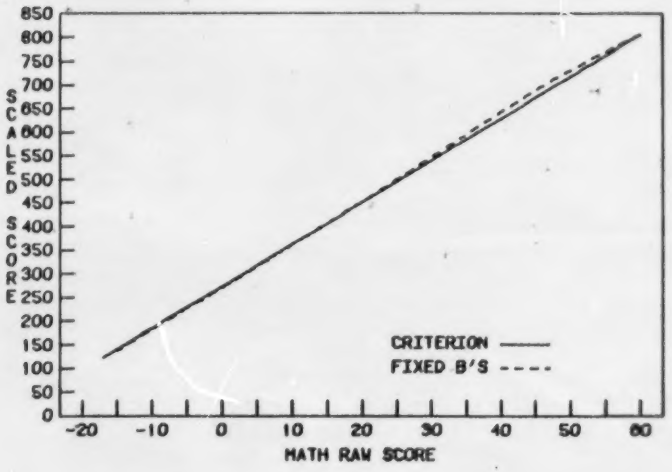


Figure 5: SAT-Mathematical chain. Comparison of initial scale (criterion) and final conversion for each equating method.





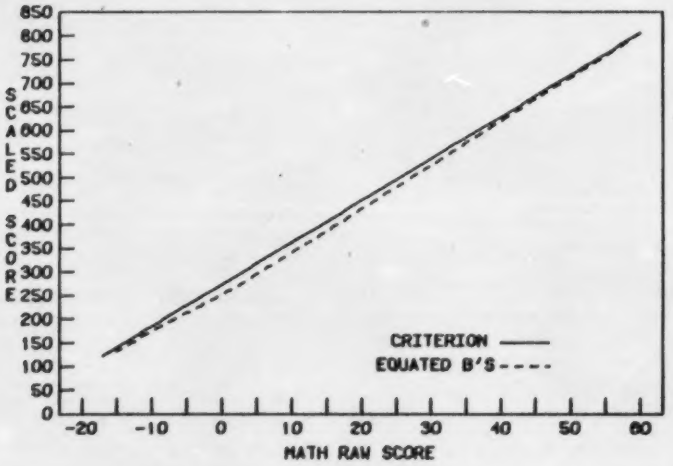


Figure 5 (continued)

to the initial scale. The linear and equipercentile conversions are the most discrepant from the initial scale.

Perusal of the SAT-Mathematical conversions shown in Figure 5 indicates that the final linear, IRT concurrent, and IRT fixed b's conversions tend to overestimate the initial scale values whereas the IRT equated b's conversion tends to underestimate the initial scale values and the equipercentile conversion tends to swing around the initial scale values. The linear, equipercentile, IRT concurrent, and IRT fixed b's conversions are similar and less discrepant than the IRT equated b's conversion from the initial scale values.

These observations are given more precise meaning by computing a discrepancy index for each comparison with the criterion. For each raw score x on form V4 there is a corresponding initial scale score t (the criterion score) and an estimated scale score t' derived from a specific equating method. The smaller the difference d between t and t', the smaller the scale drift and the more stable the equating model.

The weighted mean square difference was selected for use as a summary index to aid in the evaluation of the effectiveness of the various equating models. It was thought that the weighted mean square difference would be more useful than the mean square difference since, instead of giving equal weight to all values of x, it gives the greatest weight to those values of x which are most likely to occur. It was also thought that the weighted mean square difference would be more useful than the weighted mean absolute difference as it gives more weight to large discrepancies and is consistent with what is used to represent total error in the statistical literature.

The weighted mean square difference or total error is equal to the variance of the difference plus the squared bias, that is,

$$\sum_j f_j d_j^2/n = \sum_j f_j (d_j - \bar{d})^2/n + \bar{d}^2 \text{ , or}$$

(Total Error) = (Variance of Difference) + (Squared Bias)

where $d_j = (t'_j - t_j)$, t'_j is the estimated scale score for raw score x_j, t_j is the initial or criterion scale score for x_j, f_j is the frequency of x_j, $n = \sum_j f_j$, $\bar{d} = \sum_j f_j d_j/n$, and the summation is over that range of x for which extrapolation is unnecessary. If the ratio of the squared bias to the total error is one, then the criterion line and the conversion line are parallel. If the ratio is less than one, then there is an interaction between the criterion and the equating method.

Summary statistics and discrepancy indices for each equating method are given in Table 5. The values in Table 5 were computed from the data in Tables 3 and 4, summing over verbal raw scores 1 to 80 and mathematical raw scores -8 to 55 using the corresponding frequencies for the total group taking form V4 when it was first administered in December 1973.

Examination of the SAT-Verbal data in Table 5 indicates that the total error was largest for the equipercentile method and smallest for the IRT equated b's method. All methods, except for the IRT equated b's method, overestimated the criterion mean. The best estimate of the mean was given by the IRT equated b's method and the worst estimate by the Tucker method. All methods overestimated the criterion standard deviation. The best estimate of the standard deviation was given by the IRT equated b's method and the worst estimate by the equipercentile method. All of the methods, except for the equipercentile method, produced similar estimates

Table 5

Summary and Discrepancy Indices[a] for Equating Methods

Used in the SAT Scale Drift Study

Index	Initial Scale (Criterion)	Linear Equating			Curvilinear Equating			
						IRT Equating		
		Tucker	Levine Eq. Rel.	Levine Uneq. Rel.	Equi%ile	Con-Current	Fixed b's	Equated b's
SAT-Verbal								
Scaled Score:								
Mean	435.37	456.05	454.04	454.22	454.14	445.73	444.80	434.19
Standard Deviation	109.09	112.85	111.76	112.49	115.89	112.89	111.39	110.14
Total Error[b]		441.90	355.62	366.73	451.67	125.15	96.67	6.67
Bias		20.68	18.67	18.85	18.77	10.36	9.43	-1.18
SD of Difference		3.76	2.67	3.40	9.96	4.23	2.79	2.30
Absolute Difference:								
Maximum/Raw Score		31.61/80	26.44/80	28.73/80	44.12/63	22.19/74	17.18/66	6.02/71
Minimum/Raw Score		12.58/1	12.91/1	11.52/1	.49/4	.27/9	3.61/1	.10/12, 51
SAT-Mathematical								
Scaled Score:								
Mean	468.05	476.79	472.44	472.57	472.57	472.05	472.65	452.04
Standard Deviation	113.25	114.28	114.16	114.99	112.14	115.52	118.51	118.06
Total Error[b]		77.44	20.09	23.46	65.09	26.90	50.82	280.85
Bias		8.74	4.39	4.52	4.52	4.00	4.60	-16.02
SD of Differences		1.03	.91	1.73	6.69	3.31	5.45	4.94
Absolute Difference:								
Maximum/Raw Score		11.40/55	6.75/55	9.01/55	41.94/-5	14.48/-6	16.22/45	22.81/4
Minimum/Raw Score		6.32/-8	2.24/-8	.44/-8	.00/39	.29/3	.02/11	2.93/55

[a]Computed for SAT-Verbal raw scores 1 through 80 and for SAT-Mathematical raw scores -8 through 55.

[b]$(\text{SD of Difference})^2 + (\text{Bias})^2$. Bias equals $(M_1 - M_c)$, where M_c refers to the mean criterion score and M_1 refers to the mean yielded by a particular equating method.

for the standard deviation. Bias accounted for at least 78 percent of the total error for all methods except the IRT equated b's method for which it accounted for only 21 percent of the total error. Results were similar for the three linear models. Of the linear methods, the Levine Equally Reliable model resulted in the smallest total error and the best estimates of the criterion mean and standard deviation. All of the IRT methods produced substantially better results than the linear and equipercentile methods. Results were similar for the IRT concurrent and fixed b's methods. Of the three IRT methods, the equated b's method resulted in the smallest total error and the best estimates of the criterion mean and standard deviation. Overall, results for the IRT equated b's method seem almost too good to be true in that the results almost replicate the criterion.

Examination of the SAT-Mathematical data in Table 5 indicates that the total error was smallest for the Levine Equally Reliable model and largest for the IRT equated b's method. All methods, except for the IRT equated b's method, overestimated the criterion mean. The best estimate of the mean was given by the IRT concurrent method and the worst estimate (a substantial underestimate) by the IRT equated b's method. All methods, except for the equipercentile method, overestimated the criterion standard deviation. The best estimate of the standard deviation was given by the Levine Equally Reliable model and the worst by the IRT fixed b's method; however, the equipercentile model resulted in the largest standard deviation of the difference between the estimated and criterion scores. Of the three linear models, the Levine Equally Reliable model resulted in the smallest total error and the best estimates of the criterion mean and standard deviation; however, results were very similar for both Levine models. Of the three IRT methods, the IRT concurrent method resulted in the smallest

total error and the best estimates of the criterion mean and standard deviation. The best IRT method (concurrent) when compared to the best conventional equating method (Levine Equally Reliable) produced a better estimate of the criterion mean but a worse estimate of the criterion standard deviation. Overall, results for the two Levine models and the IRT concurrent method were similar and did not differ greatly from the criterion.

Discussion

The relationship between the linear models was similar for both the SAT-Verbal and the SAT-Mathematical equating chains. For both chains, the Tucker model resulted in the largest total error and the Levine Equally Reliable model, in the smallest total error. The Levine Equally Reliable model gave much better estimates of the criterion means than the Tucker model and slightly better estimates than the Levine Unequally Reliable model. Also, the Levine Equally Reliable model gave considerably better estimates of the criterion standard deviations than either of the other two linear models. Implicit to the derivation of the Tucker model is the assumption of random groups (Levine, 1955). Since the old and new form samples used in this study were not selected randomly from the same population, and in many instances, particularly for SAT-Verbal, differed considerably in ability level (see Table 1), it should not be surprising that the Levine models were more satisfactory than the Tucker model.

Of the conventional equating methods used in this study, equipercentile equating was less satisfactory, overall, than linear equating. The equipercentile estimates of the mean criterion scores were similar to the Levine estimates and better than the Tucker estimates. However, the standard deviations of the difference between the estimated and the

criterion scores were much larger for the equipercentile equatings than for the linear equatings. One reason that equipercentile equating may have been less satisfactory than linear equating is that, unlike the linear models, it does not take into account information contained in the bivariate distribution of total test and anchor test scores.

The IRT methods did not behave in the same manner for the verbal and mathematical equating chains. The equated b's procedure was the most satisfactory of all methods for the verbal data and the least satisfactory of all methods for the mathematical data. The concurrent method was slightly less satisfactory than the fixed b's method for the verbal data and considerably better than the fixed b's method for the mathematical data. Possibly, the IRT equated b's procedure gives less consistent results than the concurrent and fixed b's procedures because it discards data (items for which the difficulties are poorly estimated).

The relationship between the results for the conventional and IRT equating procedures is not the same for the verbal and the mathematical data. For the verbal chain, the IRT methods were greatly superior to the conventional methods. For the mathematical chain, the Levine models and the IRT concurrent method tended to give very similar results. Possibly the use of only two chains (even though the type of data differed) is too limited a sample to reject the hypothesis that all the results are due simply to random fluctuations. It is possible, however, that the mathematical data more closely approximate the underlying assumptions of the various models. Further study is needed to determine, for example, if SAT-Mathematical forms are more parallel than SAT-Verbal forms, thus resulting in less curvilinearity.

If it were necessary to use the same equating method for both SAT-Verbal and SAT-Mathematical forms, it appears that of the seven models used in this study that the IRT concurrent method would be the best equating procedure for reducing scale drift over time. This is not to say, however, that the IRT concurrent procedure would be more robust than other equating methods when applied to data on tests other than the SAT.

References

Angoff, W. H. Test reliability and effective test length. Psychometrika, 1953, 18, 1-14.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Levine, R. E. Equating the score scales of alternate forms administered to samples of different ability (RB-55-23). Princeton, N.J.: Educational Testing Service, 1955.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Erlbaum, 1980.

Mosteller, F. and Tukey, J. W. Data analysis and regression: A second course in statistics. Reading, MA: Addison and Wesley, 1977.

Wood, R L and Lord, F. M. A user's guide to LOGIST (RM-76-4). Princeton, N.J.: Educational Testing Service, 1976.

Wood, R L, Wingersky, M. S., and Lord, F. M. A computer program for estimating examinee ability and item characteristic curve parameters (RM-76-6). Princeton, N.J.: Educational Testing Service, 1976.

Appendix A

Transforming b's Using Standard Errors

Purpose of the Transformation

This procedure transforms the item parameters and ability estimates of one form (referred to as the new form) to the scale of another form (referred to as the old form). The two forms must have common items.

The transformation uses a function of the standard errors of the item difficulties (b's) for common items as weights to determine an initial transformation line based on mean and sigma equating of weighted common items. A new set of weights is computed using a combination of the standard error weights and robust (Tukey) weights based on perpendicular distances to the line. A new line is computed and the procedure iterates until the maximum change in the perpendicular distances is less than some criterion.

Method

Computing the Standard Errors

Page 191 of Lord (1980) gives the formulas for the elements of the information matrix I. This matrix is computed. The inverse of this matrix is then obtained, giving an approximation to the variance/covariance matrix among the item parameter estimates. The element on the diagonal corresponding to the item difficulty is the variance of the estimate of item difficulty. The square root of this quantity is the standard error of the estimate of item difficulty.

Each item has two item difficulties, one from each LOGIST run. Therefore, each item has two standard errors. The initial weight for an item to be used in the iterative procedure is the reciprocal of the larger squared standard error of the item difficulty.

The accuracy with which a standard error of b is computed is the ratio of the determinant to the product of the diagonals of the information matrix. If this ratio is less than 0.0001, the standard error is not accurate. The item is given a standard error weight of zero.

All people are included in the computation except those who did not reach the item.

Computing the Mean and Sigma Line Equating the Item Difficulties

Suppose we have two distributions of weighted item difficulties: BOLD(i) and BNEW(i), i = 1,..., n. We compute

BOLDM = Mean of BOLD

BOLDS = Standard Deviation of BOLD

BNEWM = Mean of BNEW

BNEWS = Standard Deviation of BNEW

The mean and sigma equating line to put the new form on the scale of the old form is

$$BOLD' = A(BNEW) + B \quad ,$$

where

BOLD' is a transformed BNEW value,

A = BOLDS/BNEWS,

B = BOLDM - A(BNEWM).

Computing the Tukey Weights

Page 205 of Mosteller and Tukey (1977) gives a method of computing a robust estimate of location by weighting data with differential weights. We use only one piece of this process, namely the formula for the weights.

For our pruposes, Y* is the transformation line we have tentatively found. We replace Tukey's (Y(i) - Y*) with the perpendicular distance of a point to the line.

Let D(i) equal the absolute value of the perpendicular distance. Then our weights, T(i), are

$$T(i) = \begin{cases} (1 - (D(i)/CS)^2)^2 & \text{when } (D(i)/CS)^2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

where S is the median of the D(i) and C is a constant equal to 6.

The Iterative Procedure

The iterative procedure proceeds as follows:

Step 1: For each item difficulty, for each common item, compute

$$W(i) = SE(B(i))^{-2},$$

where SE(B) is the larger of the two standard errors.

Step 2: Compute a vector of scaled weights

$$W(i)' = W(i)/(\text{sum of } W(i))$$

Step 3: Compute mean and sigma equating line between the two sets of b's weighted by the W' and get the slope, A, and the intercept, B.

Step 4: Compute the perpendicular distances of each point to the line.

Step 5: Compute the Tukey weights, T(i) for each item, using these perpendicular distances.

Step 6: Reweight each point by a combined weight U(i), where

$$U(i) = (W(i)*T(i))/(\text{sum of } W(i)*T(i))$$

Step 7: Compute weighted mean and sigma equating line using these new weights.

Step 8: Repeat Steps 4, 5, and 6 until the maximum change in the perpendicular distances is less than 0.01.

Result

This procedure gives low weights to poorly determined item difficulties or to item difficulties which are outliers. Once the final transformation is found for the item difficulties, the item discrimination (a's) are transformed, as well as the ability estimates.